Journal of Industrial and Management Optimization

doi:10.3934/jimo.2022076

LOCATION AND CAPACITY PLANNING FOR PREVENTIVE HEALTHCARE FACILITIES WITH CONGESTION EFFECTS

Hongzhi Lin

School of Economics and Management
Southeast University
Nanjing, China

Min Xu*

Department of Industrial and Systems Engineering
The Hong Kong Polytechnic University
Hung Hom, Hong Kong, China

Chi Xie

Key Laboratory of Road and Traffic Engineering of Ministry of Education
School of Transportation Engineering
Tongji University
Shanghai, China

(Communicated by Jung Woo Baek)

Abstract. A painful lesson got from pandemic COVID-19 is that preventive healthcare service is of utmost importance to governments since it can make massive savings on healthcare expenditure and promote the welfare of the society. Recognizing the importance of preventive healthcare, this research aims to present a methodology for designing a network of preventive healthcare facilities in order to prevent diseases early. The problem is formulated as a bilevel non-linear integer programming model. The upper level is a facility location and capacity planning problem under a limited budget, while the lower level is a user choice problem that determines the allocation of clients to facilities. A genetic algorithm (GA) is developed to solve the upper level problem and a method of successive averages (MSA) is adopted to solve the lower level problem. The model and algorithm is applied to analyze an illustrative case in the Sioux Falls transport network and a number of interesting results and managerial insights are provided. It shows that solutions to medium-scale instances can be obtained in a reasonable time and the marginal benefit of investment is decreasing.

1. Introduction.

This research is motivated by the impacts of COVID-19. The pandemic and its variants have disrupted our normal life and caused significant loss of property and life to society. It is obvious that the current healthcare system does not prepare well for serious diseases. A painful lesson got from the pandemic

2020 *Mathematics Subject Classification.* Primary: 90B22; Secondary: 90B80.

Key words and phrases. Preventive healthcare, network design, facility location, bilevel programming, user equilibrium.

*Corresponding author: Min Xu.

is that preventive healthcare is of utmost importance to governments since it can reduce the likelihood and severity of potentially life-threatening diseases by early detection. The preventive healthcare includes many services such as flu shots, vaccinations, cancer screenings, hepatitis screenings, and certainly epidemic screenings. Prevention is always better than treatment. It can save massive amounts of money for any government and improve the well-being of the people in a society. However, their uptake is not satisfying in many countries across the world.

This research concentrates on preventive healthcare network design in terms of facility location and associated capacity planning. Different from classical facility location studies where users are assigned to the closest facility, users here are recognized to choose a healthcare facility based on facility attractiveness in a competitive environment. Therefore, it is critical to understand how users make their choices. Previous studies might be divided into two types in terms of user choice behavior: (i) system optimal models, in which a central decision-maker directs where users go; (ii) user choice models, in which users are free to choose a facility. Classical facility location studies often take system optimal models with distance as the major determinant of the attractiveness of the facilities. The system optimal models still dominate the facility location studies up to date. Vidyarthi and Kuzgunkaya [17] formulated a system optimal model for the design of a preventive healthcare facility network considering waiting time at a facility. Davari et al. [2] not only took queuing limit as a constraint but also incorporated demand equity and fuzzy attractiveness by using multi-objective optimization technique. Risanger et al. [14] proposed a system optimal model to select pharmacies for COVID-19 testing where demand is an exponentially decaying function of distance. In fact, in the healthcare service industries and many other service sectors, users typically have the freedom to choose a facility to patronize, and it is more appropriate to take a user choice model (see [18]).

User choice models depicts how clients choose a facility. They can be further classified into two categories: (i) Non-equilibrium allocations, where the competition between users is not considered; (ii) Equilibrium allocations, where the competition between clients is incorporated. To be more specific, there are three ways for non-equilibrium allocation. The most popular one is all-or-nothing allocation, also known as, winner-take-all allocation, where the travel time (or distance) between user and facility is regarded as the major determinant, and clients are assumed to seek services of the closest healthcare facility in [16, 13, 4]. In addition, the effect of congestion at a facility is also recognized in recent all-or-nothing allocation. For example, Zhang et al. [21] included facility waiting time as part of total time. Davari et al. [3] and Dogan et al. [4] took waiting time as a constraint. However, these studies assumed that all users from the same node request service from the same facility with minimum time, which is not realistic. In fact, users could have more flexibility to choose facilities in practice. The second one is Huff-type allocation (see [6]), which allocates a portion of demands to a facility based on the facility's attractiveness and its travel time to the user. Note that the well-known gravity model is a special case of Huff-type allocation. The Huff-type allocation will reduce to a gravity model with a pre-specified parameter. The third one is multinomial logit allocation (see [20, 1, 5, 7]), where clients' characteristics and unobserved attributes could be included in the utility function. Although waiting time usually works as a constraint, it is not considered in facility choice decision which is not realistic. In summary, the non-equilibrium allocations dominate the facility studies as they avoid the complexity of equilibrium issues.

In contrast, more recent models adopt equilibrium allocations to incorporate the congestion effect at a facility. It has been shown by empirical analysis that the waiting time is important to users in healthcare service context [20]. However, the service waiting time is not exogenous like path travel time but endogenous. To be more specific, a shorter waiting time attracts users, but this in turn extends waiting time at a capacitated facility as more users result in congestion. This means that equilibrium issues between waiting time and user volume must be considered. Generally speaking, there are two ways for equilibrium allocations: (i) Deterministic user equilibrium allocation (see [19, 9]), where waiting time at a facility is included as an indispensable component of a deterministic utility; (ii) Stochastic user equilibrium allocation (see [18, 8, 12]), where a random component is further included to accommodate unobserved utilities. The facility with the highest (random) utility will be chosen to visit by users. These models are typically formulated as a mathematical programming with equilibrium constraints (MPEC) or a bi-level programming which is a common methodology for equilibrium problems. At equilibrium state, everyone is content with the facility they patronize, i.e., people from the same population node will have the identical utility even if they go to different facilities. The description of user choice behavior is further improved recently. For example, Kucukyazici et al. [11] adopted a latent class analysis to incorporate heterogenous user preferences in the design of cancer screening facility networks. Krohn et al. [10] further introduced the quality of healthcare into the utility function, which is defined as a dummy variable whether satisfying the minimum practitioner's quantity required. It can be expected that researches will continue towards more realistic user choice behavior.

This research makes four main theoretical and practical contributions. (i) The facility network design problem for preventive healthcare services is formulated as a bilevel programming structure. (ii) An efficient and effective heuristic solution method is proposed, which is adaptable for the analogous bilevel programming problems. (iii) The stand-alone use of the deterministic user equilibrium model to predict facility demand volumes. In particular, queuing theory is incorporated to estimate the waiting time as congestion at a facility. (iv) Several findings and managerial insights are provided based on our computational experiments.

The remainder of this article is organized as follows. Section 2 describes the problem and formulates it as a bilevel integer programming model. Section 3 proposes a heuristic algorithm to solve the bilevel programming problem. Section 4 presents computational results for the model with managerial insights. In the final section, conclusions and future research directions are presented.

2. **Problem description and model formulation.**

Consider a road network $G = (\mathbf{N}, \mathbf{L})$ with a set of nodes $\mathbf{N}$ and a set of links $\mathbf{L}$. The nodes indicate population zones or facility locations, and the links denote main transportation arteries. We assume that the number of users requiring preventive healthcare service at population node $i (i \in \mathbf{N})$ per unit of time is denoted by h_i. The set of candidate locations for healthcare facilities is $\mathbf{M} \subset \mathbf{N}$, and $\mathbf{S} \subset \mathbf{M}$ is the set of chosen facility locations. The shortest path travel time from demand node i to facility location j is denoted as t_{ij}. The government has an available investment budget B, with which one or multiple servers can be established at each chosen facility location.

We assume that servers are homogenous, and service time is distributed exponentially, providing an average number of μ clients per unit of time. We also assume that clients are homogenous, their arrivals to each facility follow Poisson distribution and the queuing discipline is first-come first-served (FCFS). These assumptions are reasonable for walk-in facilities, which apply to most routine preventive healthcare services in many countries or regions. Thus, each facility here is an $M/M/s$ queuing system, where M denotes Markovian (or Poisson) arrivals or departures distribution, or equivalently exponential interarrival or service time distribution, and s denotes the number of servers at a facility.

The goal of this study is to determine the facility location and associated capacity in order to optimize system total utility while staying within the investment budget B. First of all, three sets of decision variables are defined here:

$y_j = \begin{cases} 1 \text{ if a facility is located at location } j,\ \forall j \in \mathbf{M} \\ 0 \text{ otherwise;} \end{cases}$

s_j= number of servers at location j, $\forall j \in \mathbf{M}$;

x_{ij}= number of clients from population node i to location j, $\forall i \in \mathbf{N}$, $j \in \mathbf{M}$.

Therefore, we have $\mathbf{S} = \{j : j \in \mathbf{M}, y_j = 1\}$ and

$$\sum_{j \in \mathbf{S}} x_{ij}{=}h_i,\ \forall i \in \mathbf{N}. \tag{1}$$

Denote the arrival rate of clients at facility j by λ_j, $\forall j \in \mathbf{M}$, and we have

$$\lambda_j = \sum_{i \in \mathbf{N}} x_{ij},\ \forall j \in \mathbf{M}. \tag{2}$$

2.1. **The user utility function.** Let us now present the user choice modeling, which essentially constructs a user utility function based on the attractiveness of facilities that they are aware of. Let the observed utility of users from population node $i(i \in \mathbf{N})$ receiving the service at location j by U_{ij}. It mainly comprises three components. (i) u_j, a constant attraction of facility location j. This might include intrinsic factors such as parking convenience, facility appearance, practitioner reputation, etc. (ii) t_{ij}, the shortest path travel time from origin node i to destination facility j. (iii) $\bar{W}(\lambda_j, s_j)$, the expected waiting time at facility location $j(j \in \mathbf{M})$ including queuing time and service time, which is a function of arrival rate λ_j and server number s_j. As it is an $M/M/s_j$ queuing system at location j, for any $s_j \geq 1$, $\bar{W}(\lambda_j, s_j)$ could be given by a set of equations according to the classical queuing theory in [15]:

$$\bar{W}(\lambda_j, s_j) = \frac{L_j}{\lambda_j} + \frac{1}{\mu}\ ,\ \forall j \in \mathbf{S}, \tag{3}$$

$$L_j = \frac{\rho_j^{s_j+1}}{(s_j-1)!(s_j-\rho_j)^2} p_0,\ \forall j \in \mathbf{S}, \tag{4}$$

$$p_0 = [\sum_{n=0}^{s_j-1} \frac{\rho_j^n}{n!} + \frac{\rho_j^{s_j}}{(s_j-1)!(s_j-\rho_j)}]^{-1}\ ,\ \forall j \in \mathbf{S}, \tag{5}$$

$$\rho_j{=}\frac{\lambda_j}{\mu},\ \forall j \in \mathbf{S}, \tag{6}$$

where L_j is the expected queuing length in terms of client number, p_0 is the probability of no client, and ρ_j is the intensity of service. Note that the stability condition of the queue is assumed to be satisfied, that is, $\lambda_j < s_j\mu, \forall j \in \mathbf{S}$. Otherwise, the waiting time will be unlimited.

We adopt a conventional linear additive functional form for U_{ij} to integrate these three utility components. The other functional forms could be possible and deserved to be explored further according to latest behavior theories. It is also intuitive to assume that U_{ij} is positively associated with benefits u_j while negatively associated with costs t_{ij} and $\bar{W}(\lambda_j, s_j)$. That is, U_{ij} is formulated as

$$U_{ij}{=}u_j - \beta_1 t_{ij} - \beta_2 \bar{W}(\lambda_j, s_j), \ \forall i \in \mathbf{N}, \ j \in \mathbf{S}, \tag{7}$$

where β_1 and β_2 denote the coefficients of the travel time and the waiting time, respectively, and can be estimated empirically using actual flow data. Note that besides these specific times, the utility function can also be extended to incorporate other observed attributes, such as parking time and service quality, depending on available data.

The specific user utility is not constant but depends on the choice of other users. It is critical to note the interdependency between the arrival rate λ_j and the service waiting time $\bar{W}(\lambda_j, s_j)$. According to our modelling, λ_j is the summary of x_{ij}, which depends on U_{ij}, which further depends on $\bar{W}(\lambda_j, s_j)$ in turn. That is, the value of λ_j depends on $\bar{W}(\lambda_j, s_j)$ indirectly. The user competition will reach to an equilibrium state. It implies that we have to address a user equilibrium problem so as to determine x_{ij} and λ_j with given facility location pattern.

2.2. **The user equilibrium model.** According to utility maximization decision rule, rational clients will choose a facility with the highest observed utility. Denote $\bar{U}_i$ to be the highest utility of clients at population node i, i.e.,

$$\bar{U}_i{=}\max_{j \in \mathbf{S}} U_{ij}, \ \forall i \in \mathbf{N}. \tag{8}$$

Given location set $\mathbf{S}$ and capacities s_j, $\forall j \in \mathbf{S}$, at user equilibrium state, each client achieves his/her highest utility and he/she cannot increase his/her utility further by changing his/her choice. Therefore, the equilibrium condition can be expressed mathematically by

$$U_{ij}^* = u_j - \beta_1 t_{ij} - \beta_2 \bar{W}(\lambda_j^*, s_j) \left\{ \begin{array}{l} = \bar{U}_i^*, \text{ if } x_{ij}^* > 0 \\ \leq \bar{U}_i^*, \text{ if } x_{ij}^* = 0 \end{array} \right. , \ \forall i \in \mathbf{N}, \ j \in \mathbf{S}, \tag{9}$$

where U_{ij}^* and $\bar{U}_i^*$ represent the utility of clients at population node i that visit healthcare facility at location j and the highest utility of clients at population node i at user equilibrium state, respectively. In addition, it should be noted that

$$\lambda_j^* = \sum_{i \in N} x_{_{ij}}^*, \ \forall j \in \mathbf{S},$$

where λ_j^* denotes the arrival rate of clients at facility location j at user equilibrium sate, and $x_{_{ij}}^*$ denotes the number of clients from demand node i to facility location j at user equilibrium state.

This equilibrium condition (9) implies that if there is a client flow from demand node i to facility location j, then U_{ij}^*, the utility of clients at demand node i for facility location j, must be equal to the highest utility $\bar{U}_i^*$ that can be achieved; otherwise, it is no more than the highest. This modelling here states that each client patronizes the facility location with the highest observable attractiveness.

To find λ_j^* and implicit x_{ij}^* in Eq (9) given chosen location set $\mathbf{S}$ and associated capacities $s_j(\forall j \in \mathbf{S})$, we can solve the following equivalent nonlinear mathematical programming:

$$\max\ Z(\mathbf{x}|\mathbf{S}) = \sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{S}}\int_0^{\lambda_j} U_{ij}(\omega, s_j)d\omega \tag{10}$$

subject to

$$\sum_{j\in M} x_{ij} = h_i,\ \forall j \in \mathbf{S} \tag{11}$$

$$x_{ij} \geq 0,\ \ \forall i \in \mathbf{N},\ \ j \in \mathbf{S} \tag{12}$$

where

$$\lambda_j = \sum_{i\in N} x_{ij},\ \forall i \in \mathbf{N},\ \ j \in \mathbf{S}. \tag{13}$$

Theorem 2.1. *Given $\mathbf{S}$ and s_j, $j \in \mathbf{S}$, the mathematical programming (10)-(13) is equivalent to condition (9).*

Proof. In order to prove the mathematical programming is equivalent to Eq. (9), we reformulate it into a Lagrange function with nonnegative constraints only, i.e., □

$$\begin{aligned} &F = Z(\mathbf{x}|\mathbf{S}) - \sum_{i\in N} w_i(\sum_{j\in M} x_{ij} - h_i) \\ &s.t.\ x_{ij} \geq 0,\ \ \forall i \in \mathbf{N},\ \ j \in \mathbf{S}, \end{aligned} \tag{14}$$

where w_i in the objective function is a Lagrange multiplier to constraint (11).

According to Karush–Kuhn–Tucker (KKT) conditions, the optimal conditions of this Lagrange function are

$$x_{ij}\frac{\partial F}{\partial x_{ij}} = 0,\ \ \forall i \in \mathbf{N}, j \in \mathbf{S}, \tag{15}$$

$$\frac{\partial F}{\partial x_{ij}} \leq 0\ ,\ \ \forall i \in \mathbf{N},\ \ j \in \mathbf{S}, \tag{16}$$

$$\frac{\partial F}{\partial w_i} = 0,\ \ \forall i \in \mathbf{N}, \tag{17}$$

$$x_{ij} \geq 0,\ \ \forall i \in \mathbf{N},\ \ j \in \mathbf{S}. \tag{18}$$

It is straightforward that Eq. (17) is equivalent to Eq. (11). Eqs (15) and (16) means

$$\begin{aligned} &\text{if } x_{ij} > 0, \tfrac{\partial F}{\partial x_{ij}} = 0, \forall i \in \mathbf{N}, j \in \mathbf{S}, \\ &\text{if } x_{ij} = 0, \tfrac{\partial F}{\partial x_{ij}} \leq 0, \forall i \in \mathbf{N}, j \in \mathbf{S}. \end{aligned} \tag{19}$$

Note that,

$$\begin{aligned} &\tfrac{\partial F}{\partial x_{ij}} = \tfrac{\partial}{\partial \lambda_j}[\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{S}}\int_0^{\lambda_j} U_{ij}(\omega, s_j)d\omega]\tfrac{\partial \lambda_j}{\partial x_{ij}} - \tfrac{\partial}{\partial x_{ij}}[\sum_{i\in\mathbf{N}} w_i(\sum_{j\in\mathbf{S}} x_{ij} - h_i)] \\ &=U_{ij} - w_i. \end{aligned} \tag{20}$$

Thus, Eq. (19) can be further rewritten with Eq. (20) as follows,

$$\begin{aligned} &\text{if } x_{ij} > 0, U_{ij} - w_i = 0,\ \ \forall i \in \mathbf{N}, j \in \mathbf{S}, \\ &\text{if } x_{ij} = 0, U_{ij} - w_i \leq 0,\ \ \forall i \in \mathbf{N}, j \in \mathbf{S}. \end{aligned} \tag{21}$$

It can be also reformulated in a complementary form as follows:

$$(U_{ij} - w_i)x_{ij} = 0,\ \ \forall i \in \mathbf{N}, j \in \mathbf{S} \tag{22}$$

$$U_{ij} - w_i \leq 0,\ \ \forall i \in \mathbf{N}, j \in \mathbf{S} \tag{23}$$

$$x_{ij} \geq 0, \ \forall i \in \mathbf{N}, j \in \mathbf{S}. \tag{24}$$

Eq. (21) means that if there is demand flow $x_{ij} > 0$, the utility U_{ij} is equal to w_i; and if there is no flow, i.e., $x_{ij} = 0$, the utility U_{ij} is no more than w_i. Therefore, the Lagrange multiplier w_i can be interpreted as the highest utility $\bar{U}_i^*$ incurred by clients at population node i. Hence, Eq. (21) is the same as Eq. (9). Therefore, the solution of the mathematical programming (10)-(13) satisfy the equilibrium condition (9). In other words, we can get the equilibrium flow by solving the mathematical programming problem in Eqs. (10)-(13).

2.3. **The bilevel programming model.** The entire problem considered here is a bilevel decision structure where the upper level decision is the determination of facility locations and associated capacities, and the lower level decision is the determination of equilibrium flows of clients from demand nodes to facility locations given the upper level decisions.

There usually is a limited budget to support the establishment and operation of the preventive healthcare facilities in practice. This budget constraint can be used to incorporate the cost differences between establishing and operating facilities at different locations of an urban area. Let c_j^f be the fixed establishment cost for a facility j ($j \in \mathbf{M}$) and c^v be the unit cost of adding a server that is identical for each location. In addition, for cost effective, we assume that facilities cannot be opened unless the number of their clients exceeds a minimum workload requirement $R_{\min}$. Moreover, the number of servers at facility j cannot exceed a finite size $\hat{s}_j$.

We consider the objective of maximizing the system total utility, which is the overall observed utility of clients. The upper level model of healthcare facility network design is given by,

$$\max_{\mathbf{s}} E(\mathbf{S}) = \sum_{i \in \mathbf{N}} \sum_{j \in \mathbf{M}} x_{ij}[u_j - \beta_1 t'_{ij} - \beta_2 \bar{W}(\lambda_j, s_j)] \tag{25}$$

subject to

$$s_j \geq y_j, \ j \in \mathbf{M} \tag{26}$$

$$s_j \leq \hat{s}_j y_j, \ j \in \mathbf{M} \tag{27}$$

$$\sum_{i \in N} x_{ij} = \lambda_j, \ j \in \mathbf{M} \tag{28}$$

$$\lambda_j < s_j \mu, \ j \in \mathbf{M} \tag{29}$$

$$\lambda_j \geq R_{\min} y_j, \ j \in \mathbf{M} \tag{30}$$

$$t'_{ij} = t_{ij} + T(1 - y_j), \ i \in \mathbf{N}, \ j \in \mathbf{M} \tag{31}$$

$$\sum_{j \in M} c_j^f y_j + c^v \sum_{j \in M} s_j \leq B \tag{32}$$

$$y_j \in \{0, 1\}, \ s_j \in \text{Integer}, \ \ j \in \mathbf{M} \tag{33}$$

where x_{ij} is determined by the following lower level model given the upper level decision $\mathbf{S}$,

$$\max\ Z(\mathbf{x}|\mathbf{S}) = \sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{S}}\int_0^{\lambda_j} U_{ij}(\omega, s_j)d\omega \tag{34}$$

subject to

$$\sum_{j\in M} x_{ij} = h_i, \forall j\in\mathbf{S} \tag{35}$$

$$x_{ij}\geq 0, \forall i\in\mathbf{N}, j\in\mathbf{S}. \tag{36}$$

Objective function (25) is to maximize the total system utility. Constraint (26) ensures the assignment of at least one server to each open facility. It also ensures the nonnegativity of decision variable s_j. Constraint (27) limits the capable number of servers. Constraint (28) defines the arrival rate λ_j. Constraint (29) is the stability condition of the queues. Constraint (30) stipulates that the arrival rate at an open facility must satisfy the minimum workload requirement. Constraint (31), where T denotes a big enough number, ensures that clients only obtain the health service from open facilities. Constraint (32) is the budget and Constraints (33) define the feasible region of decision variables.

3. Solution method.

Since the bilevel programming model is highly nonlinear and contains integer decision variables, it is hard to solve exactly. Thus, we focus on efficient and effective heuristic algorithms which have many successful applications for preventive healthcare network design (see[18, 5]). The bilevel decision framework is closely followed by our proposed solution strategy. For the lower level problem, we adopt the method of successive averages (MSA) to solve the user equilibrium model. This allocation algorithm determines the equilibrium flows of clients to facilities. For the upper level problem, a meta-heuristic, generic algorithm with elitist strategy, is proposed to find the optimal locations and associated capacities. In this way, the allocation algorithm serves as an embedded module for the location algorithm. Therefore, we present the allocation algorithm first.

3.1. Allocation algorithm for the lower level model.

Given the upper level decisions $\mathbf{S}$ and s_j, $\forall j\in\mathbf{S}$, the lower level problem is to determine the equilibrium client flows. The adopted algorithm is a kind of iterative method, known as the method of successive averages (MSA). Let k be the iteration number and K be the maximum iteration number. In addition, let ε be an error tolerance parameter predetermined, and θ_k, $k=1,\ldots,K$, be a step-size parameter at iteration k with a value between zero and one. The specific computation steps are listed below.

Step 0 (Initialization): Set appropriate values for ε and K; set $k=0$; set

$$x_{ij}^0 = \frac{h_i}{|\mathbf{S}|}, \forall i\in\mathbf{N}, j\in\mathbf{S}.$$

Step 1 (Calculation of utility): Set $k := k+1$; calculate λ_j, $\forall j\in\mathbf{S}$, from Eq. (2); calculate the shortest path travel time t_{ij}, $i\in\mathbf{N}$, $j\in\mathbf{S}$, using Dijkstra's algorithm; calculate service waiting time $\bar{W}(\lambda_j, s_j)$, $j\in\mathbf{S}$, from Eqs. (3)-(6); calculate U_{ij}, $i\in\mathbf{N}$, $j\in\mathbf{S}$, according to Eq. (7); find $\bar{U}_i$, $i\in\mathbf{N}$, from Eq. (8).

Step 2 (All-or-nothing allocation): Set flow ${x'}_{ij}$ by all-or-nothing rule as follows, i.e., allocate all clients from the same population node to the most attractive facility,

$$x'_{ij} = \begin{cases} h_i, \text{ if } U_{ij} = \bar{U}_i \\ 0, \text{ if } U_{ij} < \bar{U}_i \end{cases}, \forall i \in \mathbf{N}, j \in \mathbf{S}.$$

Step 3 (Generation of search direction): Define $d_{ij} = {x'}_{ij} - x_{ij}^{k-1}$, $\forall i \in \mathbf{N}$, $j \in \mathbf{S}$, as a search direction.

Step 4 (Flow update): Update client flow $x_{ij}^k = x_{ij}^{k-1} + \theta_k d_{ij}$, $\forall i \in \mathbf{N}$, $j \in \mathbf{S}$, where θ_k is the step-size parameter given by,

$$\theta_k = \frac{1}{k+1}.$$

Step 5 (Stopping criteria): If the relative error of successive x_{ij}^k and x_{ij}^{k-1} is reached, or $k \geq K$, set $x_{ij} := x_{ij}^k$ and stop; otherwise, go to Step 1. The relative error is defined as,

$$\frac{||x_{ij}^k - x_{ij}^{k-1}||}{||x_{ij}^{k-1}||} \leq \varepsilon, \forall i \in \mathbf{N}, j \in \mathbf{S}.$$

The suggested method in each iteration determines a new search direction in Step 3 and then updates at a step-size in Step 4. The procedure continues until one of the stopping conditions in Step 5 is met. The step size θ_k in each iteration is set in advance. There could be many ways to set θ_k. In general, θ_k should decrease with k to ensure convergence. We set θ_k as the reciprocal of the iteration number $(k + 1)$. Note that there are chances that the x_{ij}^k updated in Step 4 results in a facility's arrival rate exceeding the limit allowed, which violates stability conditions (29). There usually are two ways to solve this problem: one way is to reduce the step size, and the other way is to set a large punishment time for waiting.

3.2. **Location algorithm for the upper level model.** We develop a genetic algorithm with the elitist strategy to solve the upper level problem because it is one of the most effective meta-heuristics for addressing combinatorial optimization problems, with capabilities of exploring other parts of the feasible space and avoiding local optima.

It is well-known that each chromosome represents a solution in genetic algorithms, and the quality of a solution is represented by a fitness value. In this study, an integer coding technique is employed to construct a chromosome. Each chromosome is made up of several integer numbers as genes. Each gene represents a candidate location in $\mathbf{M}$, and its value denotes the number of allocated servers. If there is no allocable server at any given location, the facility is not opened at that location. The following is how we implement the genetic algorithm with an elitist strategy:

Step 0 (Initialization): Set the used parameters, including the population size N_{pop}, the maximum number of generations Gen, the crossover probability p_c, the mutation probability p_m, the label of generation $gen = 1$, and the fraction of elitist p_e.

Step 1 (Generation of initial population): Randomly generate N_{pop} feasible solutions as an initial population of chromosomes, scattering the entire range of possible solutions. If one is not feasible according to the constraints, generate another one until a feasible solution is found.

Step 2 (Calculation of fitness function): For each chromosome in the population, the value of fitness is calculated as the objective function value. It is used to evaluate the performance of each chromosome in the population.

Step 3 (Generation of a new population):

Step 3.1 (Selection): According to the values of fitness evaluated in Step 2, the best fraction p_e is labeled for elitists, and the worst fraction p_e is discarded.

Step 3.2 (Crossover): The remaining $(1 - p_e)N_{pop}$ chromosomes are used for crossover operation. These chromosomes are matched in pairs randomly. The probability of carrying out the crossover is p_c. If the two parent chromosomes are chosen for crossover, a gene location is randomly identified to across over to generate two off-springs as new chromosomes. If newborn chromosomes are not feasible according to constraints in the upper level model, try another gene location until they are feasible.

Step 3.3 (Mutation): A chromosome is determined for mutation with probability p_m. Randomly choose two genes with at least one positive, and interchange their values. If the new chromosome is not feasible, try another two gene locations until a feasible off-spring is generated.

Step 3.4 (Elitism): Generate a new population. After genetic operations, there are still $(1 - p_e)N_{pop}$ feasible chromosomes. The labeled p_eN_{pop} elitists are added to ensure the population size N_{pop}. This allows the best chromosomes from the current generation to carry over the next generation unaltered. It guarantees that the solution quality will not decrease from one generation to the next. Update the notation of generation $gen := gen + 1$.

Step 4 (Stopping criterion): If the maximum number of generations is achieved, i.e., $gen \geq Gen$, terminate the iteration process and output the results. Otherwise, turn to Step 2.

4. **Computational experiments.**

4.1. **An illustrative case.**

We conduct computational experiments to assess the performance of proposed model and algorithm. The Sioux Falls transport network is widely used for validation in network design studies. It is a medium-sized network, as depicted in Fig.1. The network consists of 24 nodes and 76 links. In the computational experiments, it is assumed that there are 8 population nodes and 8 potential locations. Therefore, there are a total number of 64 origin-destination (O-D) pairs. The travel time and length of link a, $a \in \mathbf{L}$, denoted as t_a and l_a, respectively, are given in Table 1. The link length can be converted to the link travel time by assuming a constant link travel speed of 30 miles/hr. The healthcare demand data, in terms of the number of clients per hour (clients/hr), are listed in Table 2.

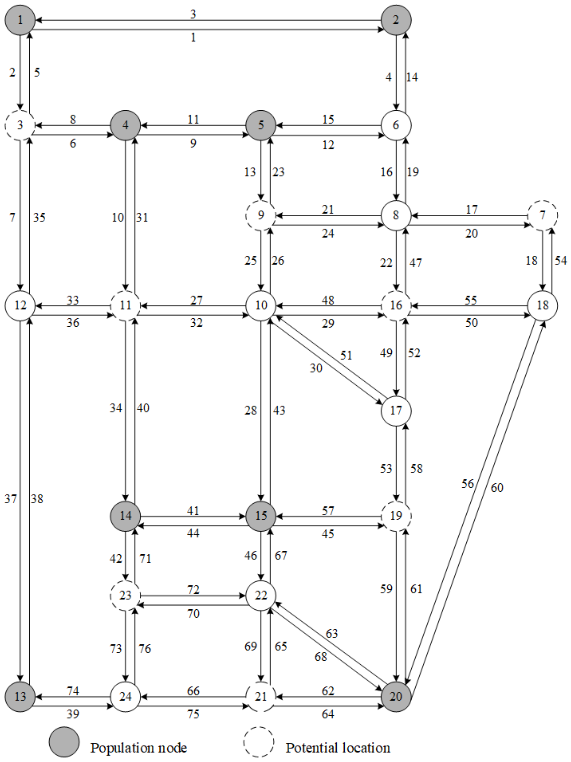

FIGURE 1. The Sioux Falls transport network

Link a	$l_a(mile)$	$t_a(h)$	Link a	$l_a(mile)$	$t_a(h)$
1,3	3.6	0.12	33,36	3.6	0.12
2,5	2.4	0.08	34,40	2.4	0.08
4,14	3	0.1	37,38	1.8	0.06
6,8	2.4	0.08	39,74	2.4	0.08
7,35	2.4	0.08	41,44	3	0.1
9,11	1.2	0.04	42,71	2.4	0.08
10,31	3.6	0.12	45,57	2.4	0.08
12,15	2.4	0.08	46,67	2.4	0.08
13,23	3	0.1	49,52	1.2	0.04
16,19	1.2	0.04	50,55	1.8	0.06
17,20	1.8	0.06	53,58	1.2	0.04
18,54	1.2	0.04	56,60	2.4	0.08
21,24	6	0.2	59,61	2.4	0.08
22,47	3	0.1	62,64	3.6	0.12
25,26	1.8	0.06	63,68	3	0.1
27,32	3	0.1	65,69	1.2	0.04
28,43	3.6	0.12	66,75	1.8	0.06
29,48	3	0.1	70,72	2.4	0.08
30,51	4.8	0.16	73,76	1.2	0.04

TABLE 1. Network characteristics for the Sioux Falls network

Population node (i)	Demandh_i(clients/hr)
1	37
2	30
4	21
5	26
13	37
14	32
15	39
20	24

TABLE 2. Healthcare demand data for the Sioux Falls network

Based on the proposed model and solution method, the following parameter values are used in the case study.

Problem parameters
the service rate of each server $\mu = 6$ client/hr;
the constant facility attraction $u_j = 0$;
the sensitivity to travel time $\beta_1 = 1$ and that to waiting time $\beta_2 = 1$;
the maximum number of servers $\hat{s}_j = 20$;
the fixed establishment cost $c_j^f = 0$;
the unit cost of a server $c^v = 1$;
the budget $B = 50$;
the minimum workload $R_{\min} = 10$ clients/hr;

Method of successive averages parameters
the maximum iteration number $K = 100$;
the error tolerance $\varepsilon = 0.01$;

Genetic algorithm parameters
the population size $N_{pop} = 200$;
the maximum number of generations $Gen = 20$;
the crossover probability $p_c = 0.5$;
the mutation probability $p_m = 0.2$;
the fraction of elitist $p_e = 0.1$.

The algorithms are coded using a free open-source language R 3.6.3. All runs are performed at a personal computer with 3.6 gigahertz Intel i7-4790 CPU and 16 gigabytes RAM. The genetic algorithm ends after 1.61 hours for this case study. The evolutionary process becomes stable after 11 generations, as shown in Fig.2. It can be concluded that the final results are satisfactory solutions. Table 3 reports the optimal results. There are four potential locations selected to set up preventive healthcare facilities, which are nodes 3, 7, 21, and 23 with associated capacities 20, 5, 13, and 12, respectively. The clients at a population node can be assigned to more than one facility such as population nodes 13 and 20. However, it also shows that clients from the same node usually patronize the same facility such as the other population nodes. Table 4 shows that the facility with the highest utility is chosen. The clients from the same population node expect to have the identical utility even if they head for different facilities. The results suggest that the utilities of visited facilities for a population node are quasi-equal.

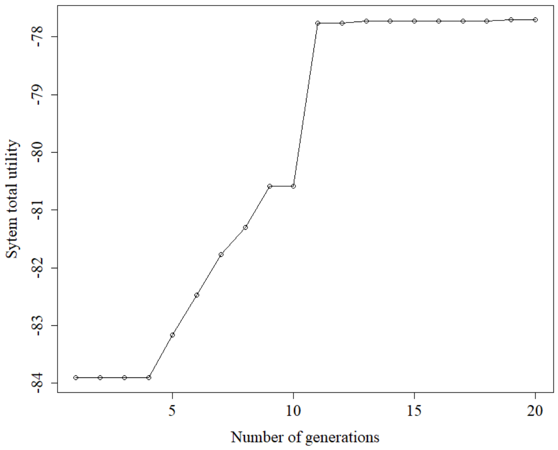

FIGURE 2. The evolutionary process of genetic algorithm

	Selected facility location (associated capacity)			
Population node	3 (20)	7 (5)	21 (13)	23 (12)
1	**36.01**	0.33	0.33	0.33
2	**19.55**	**9.91**	0.27	0.27
4	**20.44**	0.19	0.19	0.19
5	**25.3**	0.23	0.23	0.23
13	**1.65**	0.33	**6.94**	**28.08**
14	0.29	0.29	0.29	**31.14**
15	0.35	0.35	**37.96**	0.35
20	0.21	**7.07**	**16.5**	0.21

TABLE 3. Optimal network design with client flows at equilibrium state

	Selected facility location			
Population node	3	7	21	23
1	**-0.272**	-0.515	-0.550	-0.547
2	**-0.392**	**-0.395**	-0.630	-0.667
4	**-0.272**	-0.415	-0.550	-0.487
5	**-0.312**	-0.375	-0.590	-0.527
13	**-0.332**	-0.575	**-0.330**	**-0.327**
14	-0.472	-0.555	-0.370	**-0.287**
15	-0.572	-0.455	**-0.310**	-0.367
20	-0.592	**-0.315**	**-0.310**	-0.387

TABLE 4. Utility matrix between population node and selected facility location

4.2. **Sensitivity analysis.** It is always beneficial to do a sensitivity analysis which could provide valuable managerial insights. Specifically, a sensitivity analysis with varying budget control is conducted here, which is also a cost-benefit analysis. The budget is increased from 45 to 75 at step size 5. The results are shown in Fig.3, where the horizontal axis is budget, and the vertical axis is system total utility. As only travel time and waiting time are used to define utility function, the individual utility is a negative value, so as the total system utility. It is obvious that the marginal benefit is decreasing. The policymakers cannot get same benefits with same additional investments. There is an optimal budget where the marginal cost is equal to the marginal benefit.

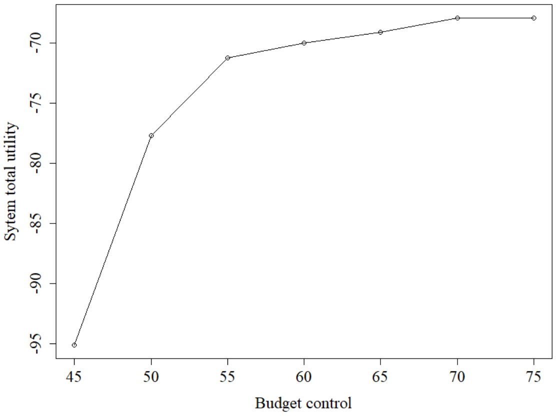

FIGURE 3. A sensitivity analysis with varying budget

As shown in Fig.3, the relationship between utility (benefit) and budget (cost) can be modeled by polynomial regression. Let f denotes the total system utility and B denotes the budget. The polynomial regression can be formulated as,

$$f(B) = \alpha_0 + \alpha_1 B + \alpha_2 B^2, \tag{37}$$

where α_0 is the intercept, α_1 is the coefficient of B, and α_2 is the coefficient of B^2. The values of these coefficients can be estimated based on the results of sensitivity analysis. The optimal budget B^* can be reached at which the marginal benefit is equal to marginal cost. That is,

$$\begin{aligned} &\tfrac{\partial f}{\partial B} = 1, \\ &\alpha_1 + 2\alpha_2 B^* = 1, \\ &B^* = \tfrac{1-\alpha_1}{2\alpha_2}. \end{aligned} \tag{38}$$

Take this sensitivity analysis as an example. The estimated parameters of Eq. (37) are shown in Table 5. The hypothesis tests show that these parameters are all significant at level 0.05. Therefore, we can reject the null hypothesis. The adjusted R-squared is 0.884, which indicates that the polynomial regression fits the data well. According to Eq. (38), the optimal budget should be 57.9. It is worthy to increase investment before the optimal budget. However, it is not wise to continue to increase investment after the optimal budget as the benefit will be less than the cost.

Coefficients	Estimate	Standard Error	t value	Pr($>|t|$))
α_0	-308.431	51.678	-5.968	0.004
α_1	7.253	1.757	4.129	0.015
α_2	-0.054	0.015	-3.718	0.021

TABLE 5. The estimated parameters for the polynomial regression

5. Conclusion.

Preventive healthcare services can improve the quality of life and make a lot of savings by diagnosing serious diseases in early stage. Governments can also utilize their healthcare expenditures more effectively, as well. This research proposes a bilevel programming model and a solution method for designing a network of preventive healthcare facilities. In the upper level model, a central decision-maker optimizes the facility location and associated capacity so as to maximize the total system utility, subject to an investment budget. In the lower level model, a user choice model addressing how clients choose a facility is adopted, where utility is defined by path travel time and service waiting time for simplicity, and the clients interact with each other to reach user equilibrium in a competitive environment. We propose a heuristic algorithm in line with the bilevel structure. A genetic algorithm with elitist strategy is proposed for the upper level model, and the method of successive averages is used for the lower level model.

To evaluate the model and the solution algorithm, we conduct a computational experiment and come up with a few noteworthy managerial insights into network design and budget control strategies. We find that the methods can reach a near-optimal solution at a reasonable time. The clients from the same node may visit more than one facility, they usually visit the same facility. A sensitivity analysis with varying budget control shows that the marginal benefit is decreasing. There is an optimal budget beyond which further increment of cost will not offset its benefit.

This research can be improved in several ways. First, since we could not find a more realistic case with available data, the well-known Sioux Falls transport network is used here as an illustrative example to show how our method and algorithm can be applied. In the future, real-life cases will be adopted to obtain more convincing results. Second, we restrict our attention to travel time and waiting time in the formulation of utility here for simplicity, but our methods could be extended directly to include other factors in the future, such as the parking time, the quality of service, the service pricing, etc. Last but not the least, the deterministic user equilibrium is adopted in this research. Future efforts could be devoted to the stochastic user equilibrium considering unobserved random utility.

Acknowledgments. This work is supported by the Key Project of Philosophy and Social Science Research in Colleges and Universities in Jiangsu Province [Grants No. 2021SJZDA025] and the Natural Science Foundation of Jiangsu Province [Grants No. BK20201277]. The second author appreciates the support from the Research Committee of The Hong Kong Polytechnic University under project code ZVPT. The third author is also supported by the National Science Foundation of China [Grant No. 72171175, 71771150, 72021002, 72111540273].

REFERENCES

[1] S. Davari, The incremental cooperative design of preventive healthcare networks, *Ann. Oper. Res.*, **272** (2019), 445–492.

[2] S. Davari, K. Kilic and G. Ertek, Fuzzy bi-objective preventive health care network design, *Health Care Management Science*, **18** (2015), 303–317.

[3] S. Davari, K. Kilic and S. Naderi, A heuristic approach to solve the preventive health care problem with budget and congestion constraints, *Appl. Math. Comput.*, **276** (2016), 442–453.

[4] K. Dogan, M. Karatas and E. Yakici, A model for locating preventive health care facilities, *Cent. Eur. J. Oper. Res.*, **28** (2020), 1091–1121.

[5] M. M. Ershadi and H. S Shemirani, Using mathematical modeling for analysis of the impact of client choice on preventive healthcare facility network design, *International J. Healthcare Management*, **14** (2021), 588–602.

[6] W. Gu, X. Wang and S. E. McGregor, Optimization of preventive health care facility locations, *International J. Health Geographics*, **9** (2010), 17.

[7] K. Haase and S. Müller, Insights into clients' choice in preventive health care facility location planning, *OR Spectrum*, **37** (2015), 273–291.

[8] M. Hamzeei and J. Luedtke, Service network design with equilibrium-driven demands, *IISE Transactions*, **50** (2018), 959–969.

[9] S. Javanmardi, H. Hosseini-nasab, A. Mostafaeipour, M. Fakhrzad and H. Khademizare, Developing a new algorithm for a utility-based network design problem with elastic demand, *International J. Engineering, Transactions B: Appl.*, **30** (2017), 758–767.

[10] R. Krohn, S. Müller and K. Haase, Preventive healthcare facility location planning with quality-conscious clients, *OR Spectrum*, **43** (2021), 59–87.

[11] B. Kucukyazici, Y. Zhang, A. Ardestani-Jaafari and L. Song, Incorporating patient preferences in the design and operation of cancer screening facility networks, *European J. Oper. Res.*, **278** (2020), 616–632.

[12] V. Marianov, M. Ríos and M. J. Icaza, Facility location for market capture when users rank facilities by shorter travel and waiting times, *European J. Oper. Res.*, **191** (2008), 32–44.

[13] M. Ndiaye and H. Alfares, Modeling health care facility location for moving population groups, *Computers and Operations Research* , **35** (2008), 2154–2161.

[14] S. Risanger, B. Singh, D. Morton and L. Meyers, Selecting pharmacies for COVID-19 testing to ensure access, *Health Care Management Science*, **24** (2021), 330–338.

[15] J. F. Shortle, J. M. Thompson, D. Gross ans C. M. Harris, *Fundamentals of Queueing Theory*, 5^{th} edition, Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 2018.

[16] V. Verter and S. D. Lapierre, Location of preventive health care facilities, *Ann. Oper. Res.*, **110** (2002), 123–132.

[17] N. Vidyarthi and O. Kuzgunkaya, The impact of directed choice on the design of preventive healthcare facility network under congestion, *Health Care Management Science*, **18** (2015), 459–474.

[18] Y. Zhang and D. Atkins, Medical facility network design: User-choice and system-optimal models, *European J. Oper. Res.*, **273** (2019), 305–319.

[19] Y. Zhang, O. Berman, P. Marcotte and V. Verter, A bilevel model for preventive healthcare facility network design with congestion, *IIE Transactions*, **42** (2010), 865–880.

[20] Y. Zhang, O. Berman and V. Verter, The impact of client choice on preventive healthcare facility network design, *OR Spectrum*, **34** (2012), 349–370.

[21] Y. Zhang, O. Berman and V. Verter, Incorporating congestion in preventive healthcare facility network design, *European J. Oper. Res.*, **198** (2009), 922–935.

Received for publication February 2022; early access May 2022.

E-mail address: linhz@seu.edu.cn
E-mail address: xumincee@gmail.com
E-mail address: chi.xie@tongji.edu.cn